DOCUMENT RESUME

ED 448 205 TM 032 235

AUTHOR Onwuegbuzie, Anthony J.

TITLE Expanding the Framework of Internal and External Validity in Quantitative Research.

PUB DATE 2000-11-21

NOTE 62p.; Paper presented at the Annual Meeting of the Association for the Advancement of Educational Research (AAER) (Ponte Vedra, FL, November 2000).

PUB TYPE Opinion Papers (120) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.

DESCRIPTORS Models; *Qualitative Research; Research Design; *Validity

ABSTRACT

An experiment is deemed to be valid, inasmuch as valid cause-effect relationships are established, if the results are due only to the manipulated independent variable (possess internal validity) and are generalizable to groups, environments, and contexts outside of the experimental settings (possess external validity). Consequently, all experimental studies should be assessed for internal and external validity. Undoubtedly, the seminal work of Donald Campbell and Julian Stanley provides the most authoritative source regarding threats to internal and external validity. Since their conceptualization, many researchers have argued that these threats to internal and external validity not only should be examined for experimental designs but are also pertinent for other quantitative research designs. Unfortunately, with respect to nonexperimental quantitative research designs, it appears that Campbell and Stanley's sources of internal and external validity do not represent the realm of pertinent threats to the validity of studies. The purpose of this paper is to provide a rationale for assessing threats to internal validity and external validity in all quantitative research studies, regardless of the research design. In addition, a more comprehensive framework of dimensions and subdimensions of internal and external validity is presented than has been undertaken previously. Different ways of expanding the discussion about threats to internal and external validity are presented. (Contains 1 figure and 58 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

Running head: FRAMEWORK FOR INTERNAL AND EXTERNAL VALIDITY

ED 448 205

Expanding the Framework of Internal and External Validity in Quantitative Research

Anthony J. Onwuegbuzie

Valdosta State University

A. J. Onwuegbuzie

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM032235

Paper presented at the annual meeting of the Association for the Advancement of Educational Research (AAER), Ponte Vedra, Florida, November, 21, 2000.

BEST COPY AVAILABLE

Abstract

An experiment is deemed to be valid, inasmuch as valid cause-effect relationships are established, if results obtained are due *only* to the manipulated independent variable (i.e., possess internal validity) and are generalizable to groups, environments, and contexts outside of the experimental settings (i.e., possess external validity). Consequently, all experimental studies should be assessed for internal and external validity. Undoubtedly the seminal work of Donald Campbell and Julian Stanley provides the most authoritative source regarding threats to internal and external validity. Since their conceptualization, many researchers have argued that these threats to internal and external validity not only should be examined for experimental designs, but are also pertinent for other quantitative research designs. Unfortunately, with respect to non-experimental quantitative research designs, it appears that the Campbell and Stanley's sources of internal and external validity do not represent the realm of pertinent threats to the validity of studies.

Thus, the purpose of the present paper is to provide a rationale for assessing threats to internal validity and external validity in *all* quantitative research studies, regardless of the research design. Additionally, a more comprehensive framework of dimensions and sub-dimensions of internal and external validity is presented than has been undertaken previously. Finally, different ways of expanding the discussion about threats to internal and external validity are presented.

Expanding the Framework of Internal and External Validity in Quantitative Research

Recently, the Committee on Professional Ethics of the American Statistical Association (ASA) addressed the following eight general topic areas relating to ethical guidelines for statistical practice: (a) professionalism; (b) responsibilities for funders, clients, and employers; (c) responsibilities in publications and testimony; (d) responsibilities to research subjects; (e) responsibilities to research team colleagues; (f) responsibilities to other statisticians or statistical practitioners; (g) responsibilities regarding allegations of misconduct; and (h) responsibilities of employers, including organizations, individuals, attorneys, or other clients utilizing statistical practitioners. With respect to *responsibilities in publications and testimony*, the Committee stated the following:

> (6) Account for all data considered in a study and explain sample(s) actually used.
>
> (7) Report the sources and assessed adequacy of the data.
>
> (8) Clearly and fully report the steps taken to guard validity.
>
> (9) Where appropriate, address potential confounding variables not included in the study. (ASA, 1999, p. 4)

Although the ASA Committee on Professional Ethics did not directly refer to these concepts, it would appear that these recommendations are related to internal and external validity.

At the same time, the ASA Committee was presenting its guidelines, the American Psychological Association (APA) Board of Scientific Affairs, who convened a committee called the Task Force on Statistical Inference, was providing recommendations for the use of statistical methods (Wilkinson & the Task Force on Statistical Inference, 1999). Useful

recommendations were furnished by the Task Force in the areas of design, population, sample, assignment (i.e., random assignment and nonrandom assignment), measurement (i.e., variables, instruments, procedure, and power and sample size), results (complications), analysis (i.e., choosing a minimally sufficient analysis, computer programs, assumptions, hypothesis tests, effect sizes, interval estimates, multiplicities, causality, tables and figures), and discussion (i.e., interpretation and conclusions).

Although the APA Task Force stated that "This report is concerned with the use of statistical methods only and is not meant as an assessment of research methods in general" (Wilkinson & the Task Force on Statistical Inference, 1999, p. 2), it is somewhat surprising that internal and external validity was mentioned directly only once. Specifically, when discussing the reporting of instruments, the task force declared:

> There are many methods for constructing instruments and psychometrically validating scores from such measures. Traditional true-score theory and item-response test theory provide appropriate frameworks for assessing reliability and *internal validity*. Signal detection theory and various coefficients of association can be used to assess *external validity*. [emphasis added] (p. 5)

The APA Task Force also stated (a) "In the absence of randomization, we should do our best to investigate sensitivity to various untestable assumptions" (p. 4); (b) "Describe any anticipated sources of attrition due to noncompliance, dropout, death, or other factors" (p. 6); (c) "Describe the specific methods used to deal with experimenter bias, especially if you collected the data yourself" (p. 4); (d) "When you interpret effects, think of credibility, generalizability, and robustness" (p. 16) ; (e) "Are the design and analytic methods robust

enough to support strong conclusions?" (p. 16); and (f) "Remember, however, that acknowledging limitations is for the purpose of qualifying results and avoiding pitfalls in future research" (p. 16). It could be argued that these six statements pertain to validity. However, the fact that internal and external validity was not directly mentioned by the ASA Committee on Professional Ethics, as well as the fact that these concepts were mentioned only once by the APA Task Force and were not directly referenced in the "Discussion" section of the report, is a cause for concern, bearing in mind that the issue of internal and external validity not only is regarded by instructors of research methodology, statistics, and measurement as being the most important in their fields, but that it also receives the most extensive coverage in their classes (Mundfrom, Shaw, Thomas, Young, & Moore, 1998).

In experimental research, the researcher manipulates at least one independent variable (i.e., the hypothesized cause), attempts to control potentially extraneous (i.e., confounding) variables, and then measures the effect(s) on one or more dependent variables. According to quantitative research methodologists, experimental research is the only type of research in which hypotheses concerning cause-and-effect relationships can be validly tested. As such, proponents of experimental research believe that this design represents the apex of research. An experiment is deemed to be valid, inasmuch as valid cause-effect relationships are established, if results obtained are due *only* to the manipulated independent variable (i.e., possess internal validity) and are generalizable to groups, environments, and contexts outside of the experimental settings (i.e., possess external validity). Consequently, according to this conceptualization, all experimental studies should be assessed for internal and external validity.

A definition of internal validity and external validity can be found in any standard research methodology textbook. For example, Gay and Airasian (2000, p. 345) describe internal validity as "the condition that observed differences on the dependent variable are a direct result of the independent variable, not some other variable." As such, internal validity is threatened when plausible rival explanations cannot be eliminated. Johnson and Christensen (2000, p. 200) define external validity as "the extent to which the results of a study can be generalized to and across populations, settings, and times." Even if a particular finding has high internal validity, this does not mean that it can be generalized outside the study context.

Undoubtedly the seminal works of Donald Campbell and Julian Stanley (Campbell, 1957; Campbell & Stanley, 1963) provide the most authoritative source regarding threats to internal and external validity. Campbell and Stanley identified the following eight threats to internal validity: history, maturation, testing, instrumentation, statistical regression, differential selection of participants, mortality, and interaction effects (e.g., selection-maturation interaction) (Gay & Airasian, 2000). Additionally, building on the work of Campbell and Stanley, Smith and Glass (1987) classified threats to external validity into the following three areas: population validity (i.e., selection-treatment interaction), ecological validity (i.e., experimenter effects, multiple-treatment interference, reactive arrangements, time and treatment interaction, history and treatment interaction), and external validity of operations (i.e., specificity of variables, pretest sensitization).

Although experimental research designs are utilized frequently in the physical sciences, this type of design is not as commonly used in social science research in general

and educational research in particular due to the focus on the social world as opposed to the physical world. Nevertheless, since Campbell and Stanley's conceptualization, some researchers (e.g., Huck & Sandler, 1979; McMillan, 2000) have argued that threats to internal and external validity not only should be evaluated for experimental designs, but are also pertinent for other types of quantitative research (e.g., descriptive, correlational, causal-comparative, quasi-experimental). Unfortunately, with respect to non-experimental quantitative research designs, it appears that the above sources of internal and external validity do not represent the realm of pertinent threats to the validity of studies.

Thus, the purpose of the present paper is to provide a rationale for assessing threats to internal validity and external validity in *all* quantitative research studies, regardless of the research design. After providing this rationale, the discussion will focus on providing additional sources of internal and external validity. In particular, a more comprehensive framework of dimensions and sub-dimensions of internal and external validity will be presented than has been undertaken previously. Brief heuristic examples will be given for each of these new dimensions and sub-dimensions. Finally, different ways of expanding the discussion about threats to internal and external validity will be presented.

UTILITY OF DELINEATING THREATS TO INTERNAL AND EXTERNAL VALIDITY

Despite the recommendations of the ASA Committee on Professional Ethics (ASA, 1999) and the APA Task Force (Wilkinson & the Task Force on Statistical Inference, 1999), a paucity of researchers provide a commentary of threats to internal and external validity in the discussion section of their articles. Onwuegbuzie (2000a) reviewed the prevalence of discussion of threats to internal and external validity in empirical research reports

published in several reputable journals over the last few years, including the American Educational Research Journal (AERJ)--a flagship journal. With respect to the AERJ, Onwuegbuzie found that although 5 (31.3%) of the 16 quantitative-based research articles published in 1998 contained a general statement in the discussion section that the findings had limited generalizability, only 1 study utilized the term "external validity." The picture regarding internal validity was even more disturbing, with none of the 16 articles published that year containing a discussion of any threats to internal validity. Moreover, in almost all of these investigations, implications of the findings were discussed as if no rival hypotheses existed. In many instances, this may give the impression that confirmation bias took place, in which theory confirmation was utilized instead of theory testing (Greenwald, Pratkanis, Leippe, & Baumgardner, 1986).

As stated by Onwuegbuzie (2000a), authors' general failure to discuss threats to validity likely stems from a fear that to do so would expose any weaknesses in their research, which, in turn, might lead to their manuscripts being rejected by journal reviewers. Yet, it is clear that every single study in the field of education has threats to internal and external validity. For example, instrumentation can never be fully eliminated as a potential threat to internal validity because outcome measures can never yield scores that are perfectly reliable or valid. Thus, whether or not instrumentation is acknowledged in a research report, does not prevent it from being a validity threat. With respect to external validity, all samples, whether random or non-random are subject to sampling error. Thus, population and ecological validity is a threat to external validity in virtually all educational studies.

The fact that the majority of empirical investigations do not contain a discussion of threats to internal and external validity also probably stems from a misperception on the part of some researchers that such threats are only relevant in experimental studies. For other researchers, failure to mention sources of invalidity may arise from an uncompromising positivistic stance. As noted by Onwuegbuzie (2000b), pure positivists contend that statistical techniques are objective; however, they overlook many subjective decisions that are made throughout the research process (e.g., using a 5% level of significance). Further, the lack of random sampling prevalent in educational research, which limits generalizability, as well as the fact that variables can explain as little as 2% of the variance of an outcome measure to be considered non-trivial, make it clear that all empirical research in the field of education are subject to considerable error. This should prevent researchers from being as adamant about the existence of positivism in the social sciences as in the physical sciences (Onwuegbuzie, 2000b).

Moreover, discussing threats to internal and external validity has at least three advantages. First and foremost, providing information about sources of invalidity allows the reader to place the researchers' findings in their proper context. Indeed, failure to discuss the limitations of a study may provide the reader with the false impression that no external replications are needed. Yet, replications are the essence of research (Onwuegbuzie & Daniel, 2000; Thompson, 1994a). Second, identifying threats to internal and external validity helps to provide directions for future research. That is, replication studies can be designed to minimize one or more of these validity threats identified by the researcher(s).

Third, once discussion of internal and external validity becomes commonplace in

research reports, *validity meta analyses* could be conducted to determine the most prevalent threats to internal and external validity for a given research hypothesis. These *validity meta analyses* would provide an effective supplement to traditional meta analyses. In fact, *validity meta analyses* could lead to *thematic* effect sizes being computed for the percentage of occasions in which a particular threat to internal or external validity is identified in replication studies (Onwuegbuzie, 2000c). For example, a narrative that combines traditional meta analyses and *validity meta analyses* could take the following form:

> Across studies, students who received Intervention A performed on standardized achievement tests, on average, nearly two-thirds of a standard deviation (Cohen's (1988) *Mean d* = .65) higher than did those who received Intervention B. This represents a moderate-to-large effect. However, these findings are tempered by the fact that in these investigations, several threats to internal validity were noted. Specifically, across these studies, *statistical regression* was the most frequently identified threat to internal validity (prevalence rate/effect size = 33%), followed by *mortality* (effect size = 22%). With respect to external validity, *population validity* was the most frequently cited threat (effect size = 42%), followed by reactive arrangements (effect size = 15%)....

Such *validity meta analyses* would help to promote the use of external replications and to minimize the view held by some researchers that a single carefully-designed study could serve as a panacea for solving educational problems (Onwuegbuzie, 2000c).

FRAMEWORK FOR IDENTIFYING THREATS TO INTERNAL AND EXTERNAL VALIDITY

As noted by McMillan (2000), threats to internal and external validity typically are presented with respect to experimental research designs. Consequently, most authors of research methodology textbooks tend to present only the original categories of validity threats conceptualized by Campbell and Stanley (1963). Unfortunately, this framework does not represent the range of validity threats. Thus, it is clear that in order to promote the discussion of threats to internal and external validity in all empirical reports, regardless of research design used, Campbell and Stanley's framework needs to be expanded. Without such an expansion, for example, many threats to internal validity outside these threats will continue to be labeled as *history*.

Surprisingly, despite Huck and Sandler's (1979) recommendation that researchers extend the classic list of seven threats to internal validity identified by Campbell and Stanley, an extensive review of the literature revealed only two articles representing a notable attempt to expand Campbell and Stanley's framework. Specifically, Huck and Sandler (1979) presented 20 categories of threats to validity, which they termed rival hypotheses. Unfortunately, using this label gives the impression that threats to internal and external validity are only pertinent for empirical studies in which hypotheses are tested. Yet, these threats also are pertinent when descriptive research designs are utilized. For example, in research in which no inferences are made (e.g., descriptive survey research), instrumentation typically is a threat to internal validity inasmuch as if the survey instrument does not lead to valid responses, then descriptive statistics that arise from the survey responses, however simple, will be invalid.

Thus, Huck and Sandler's (1979) list although extremely useful, falls short of providing a

framework that is applicable for all empirical research.

More recently, McMillan (2000) presented a list of 54 threats to validity. Moreover, McMillan re-named internal validity as internal credibility, which he defined as "rival explanations to the propositions made on the basis of the data" (p. 2). McMillan further subdivided his threats to internal credibility into three categories, which he labeled (a) statistical conclusion, (b), relationship conclusion, and (c) causal conclusion. According to this theorist, statistical conclusion threats are threats that are statistically based (e.g., small effect size); relationship conclusion threats are mostly related to correlational and quasi-experimental research designs; and causal conclusion mostly pertain to experimental research designs. McMillan (2000) also renamed external validity as generalizability in an attempt to provide a more "conceptually clear and straightforward" definition (p. 3). He divided threats that fall into these categories as population validity and ecological validity.

In short, McMillan produced a 2 (experimental vs. nonexperimental) x 3 (statistical conclusion, relationship conclusion, causal conclusion) matrix for internal credibility, and a 2 (experimental vs. nonexperimental) x 2 (population validity vs. ecological validity) matrix for generalizability. Perhaps the most useful aspect of McMillan's re-conceptualization of Campbell and Stanley's threats to internal and external validity is the fact that threats were categorized as falling into either an experimental or non-experimental design. However, as is the case for Huck and Sandler's (1979) conceptualization, McMillan's two matrices is still not as integrative with respect to quantitative research designs as perhaps it could be.

Thus, what follows is a re-conceptualization of Campbell and Stanley's (1963)

threats to internal and external validity, which further builds on the work of Huck and Sandler (1979) and McMillan (2000). Interestingly, threats to internal validity and external validity can be renamed as threats to *internal replication* and *external replication*, respectively. An internal replication threat represents the extent to which the results of a study would re-occur if the study was replicated using exactly the same sample, setting, context, and time. If the independent variable truly was responsible for changes in the dependent variable, with no plausible rival hypotheses, then conducting an internal replication of the study would yield exactly the same results. On the other hand, an external replication threat refers to the degree that the findings of a study would replicate across different populations of persons, settings, contexts, and times. If the sample was truly generalizable, then external replications across different samples would produce the same findings. However, rather than labeling these threats internal replication and external replication, as was undertaken by Huck and Sandler (i.e., rival hypotheses) and McMillan (i.e., internal credibility and generalizability), for the purposes of the present re-conceptualization, the terms internal validity and external validity were retained. It was believed that keeping the original labels would reduce the chances of confusion especially among graduate students and, at the same time, increase the opportunity that this latest framework will be diffused (Rogers, 1995). Further, the reader will notice that rather than use the term experimental group, which connotes experimental designs, the term *intervention* group has been used, which more accurately reflects the school context, whereby interventions typically are implemented in a non-randomized manner.

Threats to internal and external validity can be viewed as occurring at one or more

of the three major stages of the inquiry process, namely: research design/data collection, data analysis, and data interpretation. Unlike the case for qualitative research, in quantitative research, these stages typically represent three distinct time points in the research process. Figure 1 presents a concept map of the major dimensions of threats to internal and external validity at the three major stages of the research process. What follows is a brief discussion of each of the threats to validity dimensions and their subdimensions.

Insert Figure 1 about here

Research Design/Data Collection

Threats to Internal Validity

As illustrated in Figure 1, the following 22 threats to internal validity occur at the research design/data collection stage. These threats include Campbell and Stanley's (1963) 8 threats to internal validity, plus an additional 14 threats.

History. This threat to internal validity refers to the occurrence of events or conditions that are unrelated to the treatment but that occur at some point during the study to produce changes in the outcome measure. The longer an inquiry lasts, the more likely that history will pose a threat to validity. History can stem from either internal or extraneous events. With respect to the latter, suppose that counselors and teachers in a high school conducted a series of workshops for all students that promoted multiculturism and diversity. However, suppose that between the time that the series of workshops ended and the post-

intervention outcome measure (e.g., attitudes toward ethnic integration) was administered, a racial incident took place in a nearby school that received widespread media coverage. Such an occurrence could easily reduce the effectiveness of the workshop and thus threaten internal validity by providing rival explanations of subsequent findings.

Maturation. Maturation pertains to the processes that operate within a study participant due, at least in part, to the passage of time. These processes lead to physical, mental, emotional, and intellectual changes such as aging, boredom, fatigue, motivation, and learning, that can be incorrectly attributed to the independent variable. Maturation is particularly a concern for younger study participants, such as Kindergartners.

Testing. Testing, also known as *pretesting* and *pretest sensitization*, also refers to changes that may occur in participants' scores obtained on the second administration or post-intervention measure as a result of having taken the pre-intervention instrument. In other words, being administered a pre-intervention instrument may improve scores on the post-intervention measure regardless of whether any intervention takes place in between. Testing is more likely to prevail when (a) cognitive measures are utilized that involve the recall of factual information and (b) the time between administration is short. When cognitive tests are administered, a pre-intervention measure may lead to increased scores on the post-intervention measure because the participants are more familiar with the testing format and condition, have developed a strategy for increasing performance, are less anxious about the test on the second occasion, or can remember some of their prior responses and thus make subsequent adjustments. With attitudes and measures of personality and other affective variables, being administered a pre-intervention measure

may induce participants subsequently to reflect about the questions and issues raised during the pre-intervention administration and to supply similar or different responses to the post-intervention measure as a result of this reflection.

Instrumentation. The Instrumentation threat to internal validity occurs when scores yielded from a measure lacks the appropriate level of consistency (i.e., low reliability) or does not generate valid scores, as a result of inadequate content-, criterion-, and/or construct-related validity. Instrumentation can occur in many ways, including when (a) the post-intervention measure is not parallel (e.g., different level of difficulty) to the pre-intervention measure (i.e., the test has low equivalent-forms reliability); (b) the pre-intervention instrument leads to unstable scores regardless of whether or not an intervention takes place (i.e., has low test-retest reliability); (c) at least one of the measures utilized does not generate reliable scores (i.e., low internal-consistency reliability; and (d) the data are collected through observation, and the observing or scoring is not consistent from one situation to the next within an observer (i.e., low *intra-rater reliability*) or is not consistent among two or more data collectors/analysts (i.e., low *inter-rater reliability*).

Statistical regression. Statistical regression typically occurs when participants are selected on the basis of their extremely low or extremely high scores on some pre-intervention measure. This phenomenon refers to the tendency for extreme scores to regress, or move toward, the mean on subsequent measures. Interestingly, many educational researchers study special groups of individuals such as at-risk children with learning difficulties or disabilities. These special populations usually have been identified because of their extreme scores on some outcome measure. A researcher often cannot

be certain whether any post-intervention differences observed for these individuals are real or whether they represent statistical artifacts. According to Campbell and Kenny (1999), regression toward the mean is an artifact that can be due to extreme group selection, matching, statistical equating, change scores, time-series studies, and longitudinal studies. Thus, statistical regression is a common threat to internal validity in educational research.

Differential selection of participants. Differential selection of participants, also known as selection bias, refers to substantive differences between two or more of the comparison groups prior to the implementation of the intervention. This threat to internal validity, which clearly becomes realized at the data collection stage, most often occurs when already-formed (i.e., non-randomized) groups are compared. Group differences may occur with respect to cognitive, affective, personality, or demographic variables. Unfortunately, it is more difficult to conduct controlled, randomized studies in natural educational settings, thus differential selection of participants is a common threat to internal validity. Thus, investigators always should strive to assess the equivalency of groups by comparing groups with respect to as many variables as possible. Indeed, such equivalency checks should be undertaken even when randomization takes place, because although randomization increases the chances of group equivalency on important variables, it does not guarantee this equality. That is, regardless of the research design, when groups are compared, selection bias always exists to some degree. The greater this bias, the greater the threat to internal validity.

Mortality. Mortality, also known as attrition, refers to the situation in which participants who have been selected to participate in a research study either fail to take

part at all or do not participate in every phase of the investigation (i.e., drop out of the study). However, a loss of participants, per se, does not necessarily produce a bias. This bias occurs when participant attrition leads to differences between the groups that cannot be attributed to the intervention. Mortality-induced discrepancy among groups often eventuates when there is a differential loss of participants from the various treatment conditions, such that an inequity develops or is exacerbated on variables other than the independent variable. Mortality often is a threat to internal validity when studying at-risk students who tend to have lower levels of persistence, when volunteers are utilized in an inquiry, or when a researcher is comparing a new intervention to an existing method.

Because dropping out of a study is often the result of relatively low levels of motivation, persistence, and the like, a greater loss in attrition in the control group may attenuate any true differences between the control and intervention groups due to the fact that the control group members who remain are closer to the intervention group members with respect to these affective variables. Conversely, when a greater attrition rate occurs in the intervention group, group differences measured at the end of the study period may be artificially inflated because individuals who remain in the inquiry represent more motivated or persistent members. Both these scenarios provide rival explanations of observed findings. In any case, the researcher should never assume that mortality occurs in a random manner and should, whenever possible, (a) design a study that minimizes the chances of attrition; (b) compare individuals who withdraw from the investigation to those who remain, with respect to as many available cognitive, affective, personality, and demographic variables as possible; and (c) attempt to determine the precise reason for

withdrawal for each person.

Selection interaction effects. Many of the threats to internal validity presented above also can interact with the differential selection of participants to produce an effect that resembles the intervention effect. For example, a *selection by mortality* threat can occur if one group has a higher rate of attrition than do the other groups, such that discrepancies between groups create factors unrelated to the intervention that are greater as a result of differential attrition than prior to the start of the investigation. Similarly, a *selection by history* threat would occur if individuals in the groups experience different history events, and that these events differentially affected their responses to the intervention. A *selection by maturation* interaction would occur when one group has a higher rate of maturation than do the other groups (even if no pretest differences prevailed), and that this higher rate accounts for at least a portion of the observed effect. This type of interaction is common when volunteers are compared to non-volunteers.

Implementation bias. Although not a threat recognized by Campbell and Stanley (1963), McMillan (2000), nor Huck and Sandler (1979), implementation bias is a common and serious threat to internal validity in many educational intervention studies. Indeed, it is likely that implementation bias is the most frequent and pervasive threat to internal validity at the data collection stage in intervention studies. Implementation bias often stems from *differential selection of teachers* who apply the innovation to the intervention groups. In particular, as the number of instructors involved in an instructional innovation increases, so does the likelihood that at least some of the teachers will not implement the initiative to its fullest extent. Such lack of adherence to protocol on the part of some teachers might

stem from lack of motivation, time, training, or resources; inadequate knowledge or ability; poor self-efficacy; implementation anxiety; stubbornness; or poor attitudes. Whatever the source, implementation bias leads to the protocol designed for the intervention not being followed in the intended manner (i.e., *protocol bias*). For example, poor attitudes of some of the teachers toward an innovation may lead to intervention protocol being violated, which then transgresses to their students, resulting in effect sizes being attenuated. A particularly common component of the implementation threat that prevails is related to time. Many studies involve the assessment of an innovation after one year or even less, which often is an insufficient time frame to observe positive gains. Differences in teaching experience between teachers participating in the intervention and non-intervention groups is another way in which implementation bias may pose a threat to internal validity.

Sample augmentation bias. Sample augmentation bias is another threat to internal validity that does not appear to have been mentioned formally in the literature. This form of bias, which essentially is the opposite of mortality, prevails when one or more individuals join the intervention or non-intervention groups. In the school context, this typically happens when students (a) move away from a school that is involved in the study, (b) move to a school involved in the research from a school that was not involved in the investigation, or (c) move from an intervention school to a non-intervention school. In each of these cases, not all students receive the intervention for the complete duration of the study. Thus, sample augmentation bias can either increase or attenuate the effect size.

Behavior bias. Also, not presented in the literature, is behavior bias that occurs when an individual has a strong personal bias in favor of or against the intervention prior

to the beginning of the study. Such a bias would lead to a protocol bias that threatens internal validity. Behavior bias is most often a threat when participants are exposed to all levels of a treatment.

Order bias. When multiple interventions are being compared in a research study, such that all participants are exposed to and measured under each and every intervention condition, an order effect can provide a threat to internal validity when the effect of the order of the intervention conditions cannot be separated from the effect of the intervention conditions. For example, any observed differences between the intervention conditions may actually be the result of a practice effect or a fatigue effect. Further, individuals may succumb to the primacy or recency effect. Thus, in these types of studies, researchers should vary the order in which the intervention was presented, preferably in a random manner (i.e., counterbalancing).

Observational bias. Observational bias occurs when the data collectors have obtained an insufficient sampling of the behavior(s) of interest. This lack of adequate sampling of behaviors happens if either persistent observation or prolonged engagement does not occur (Lincoln & Guba, 1985).

Researcher bias. Researcher bias may occur during the data collection stage when the researcher has a personal bias in favor of one technique over another. This bias may be subconsciously transferred to the participants in such a way that their behavior is affected. In addition to influencing the behavior of participants, the researcher could affect study procedures or even contaminate data collection techniques. Researcher bias particularly is a threat to internal validity when the researcher also serves as the person

implementing the intervention. For example, if a teacher-researcher investigating the effectiveness of a new instructional technique that he or she has developed and believes to be superior to existing strategies, he or she may unintentionally influence the outcome of the investigation.

Researcher bias can be either *active* or *passive*. Passive sources include personality traits or attributes of the researcher (e.g., gender, ethnicity, age, type of clothing worn), whereas active sources may include mannerisms and statements made by the researcher that provide an indication of the researcher's preferences. Another form of researcher bias is when the researcher's prior knowledge of the participants differentially affects the participants' behavior. In any case, the optimal approach to minimize this threat to internal validity is to let other trained individuals, rather than the researcher, work directly with the study participants, and perhaps even collect the data.

Matching bias. A researcher may use matching techniques to select a series of groups of individuals (e.g., pairs) who are similar with respect to one or more characteristics, and then assign each person within each group to one of the treatment conditions. Alternatively, once participants have been selected for one of the treatment conditions, a researcher may find matches for each member of this condition and assign these matched individuals to the other treatment group(s). Unfortunately, this poses a threat to internal validity in much the same way as does the mortality threat. Specifically, because those individuals from the sampling frame for whom a match cannot be found are excluded from the study, any difference between those selected and those excluded may lead to a statistical artifact. Indeed, even though matching eliminates the possibility that the

independent variable will be confounded with group differences on the matching variable(s), a possibility remains, however, that one or more of the variables not used to match the groups may be more related to the observed findings than is the independent variable.

Treatment replication error. Treatment replication error occurs when researchers collect data that do not reflect the correct unit of analysis. The most common form of treatment replication error is when an intervention is administered once to each group of participants or to a few classes or other existing groups, yet only individual outcome data are collected (McMillan, 1999). As eloquently noted by McMillan (1999), such practice seriously violates the assumption that each replication of the intervention for each and every participant is independent of the replications of the intervention for all other participants. If there is one administration of the intervention to a group, whatever peculiarities that prevail as a result of that administration are confounded with the intervention. Moreover, systematic error likely ensues (McMillan, 1999). Further, individuals within a group likely influence one another in a group context when being measured by the outcome measure(s). Such confounding provides rival explanations to any subsequent observed finding, thereby threatening internal validity at the data collection stage. This confounding is even more severe when the intervention is administered to groups over a long period of time because the number of confounding variables increases as a function of time (McMillan, 1999). Both McMillan (1999) and Onwuegbuzie and Collins (2000) noted that the majority of the research in the area of cooperative learning is flawed because of this treatment replication error.

Disturbingly, treatment replication errors also occur in the presence of randomization of participants to groups, specifically when the intervention is assigned to and undertaken in groups--that is, when each participant does not respond independently from other participants (McMillan, 1999). Thus, researchers should collect data at the group level for subsequent analysis when there is a limited number of interventions independently replicated.

Evaluation anxiety. There is little doubt that in the field of education, achievement is the most common outcome measure. Unfortunately, evaluation anxiety, which is experienced by many students, has the potential to threaten internal validity by introducing systematic error into the measurement. This threat to internal validity stemming from evaluation anxiety occurs at all levels of the educational process. For example, Onwuegbuzie and Seaman (1995) found that graduate students with high levels of statistics test anxiety who were randomly assigned to a statistics examination that was administered under timed conditions tended to have lower levels of performance than did their high-anxious counterparts who were administered the same test under untimed conditions. These researchers concluded that when timed examinations are administered, the subsequent results may be more reflective of anxiety level than of actual ability or learning that has taken place as the result of the intervention. Similarly, at the elementary and secondary school level, Hill and Wigfield (1984) have suggested that examination scores of students with high levels of test anxiety, obtained under timed examination conditions, may represent an invalid lower-bound estimate of their actual ability or aptitude. Thus, researchers should be cognizant of the potential confounding role of the testing

environment at the research design/data collection stage.

Multiple-treatment interference. Multiple-treatment interference occurs when the same research participants are exposed to more than one intervention. Multiple-treatment interference exclusively (e.g., Campbell & Stanley, 1963) has been conceptualized as a threat to external validity. However, this interference also threatens internal validity. Specifically, when individuals receive multiple interventions, carryover effects from an earlier intervention may make it difficult to assess the effectiveness of a later treatment, thereby providing rival explanations of the findings. Thus, a sufficient *washout* period is needed for the effects of the previous intervention to dissipate, if this is possible. Typically, the less time that elapses between the administration of the interventions, the greater the threat to internal validity. Therefore, when designing studies in which participants receive multiple interventions, researchers should seek to maximize the washout period, as well as to counterbalance the administration of the interventions.

Reactive arrangements. Reactive arrangements, also known as *reactivity* or *participant effects*, refer to a number of facets related to the way in which a study is undertaken and the reactions of the participants involved. In other words, reactive arrangements pertain to changes in individuals' responses that can occur as a direct result of being aware that one is participating in a research investigation. For example, the mere presence of observers or equipment during a study may alter the typical responses of students that rival explanations for the findings prevail, which, in turn, threaten internal validity. In virtually all research methodology textbooks, reactive arrangements is labeled solely as a threat to external validity. Yet, reactive arrangements also provide a threat to

internal validity by confounding the findings and providing rival explanations.

Reactive arrangements comprise the following five major components: (a) the *Hawthorne effect*, (b) the *John Henry effect*, (c) *resentful demoralization*, (d) the *novelty effect*, and (e) the *placebo effect*. The Hawthorne effect represents the situation when individuals interpret their receiving an intervention as being given *special* attention. As such, the participants' reaction to their perceived special treatment is confounded with the effects of the intervention. The Hawthorne effect tends to increase the effect size because individuals who perceive they are receiving preferential treatment are more likely to participate actively in the intervention condition.

The John Henry effect, or *compensatory rivalry*, occurs when on being informed that they will be in the control group, individuals selected for this condition decide to compete with the new innovation by expending extra effort during the investigation period. Thus, the John Henry effect tends to reduce the effect size by artificially increasing the performance of the control group. Resentful demoralization is similar to the John Henry effect inasmuch as it involves the reaction of the control group members. However, instead of knowledge of being in the control group increasing their performance levels, they become resentful about not receiving the intervention, interpret this as a sign of being ignored or disregarded, and become demoralized. This loss of morale consequently leads to a reduction in effort expended and subsequent decrements in performance or other outcomes. Thus, resentful demoralization tends to increase the effect size.

The novelty effect, refers to increased motivation, interest, or participation on the part of study participants merely because they are undertaking a different or novel task.

The novelty effect is a threat to internal validity because it competes with the effects of the intervention as an explanation to observed findings. Unlike the Hawthorne, John Henry, and resentful demoralization effects, in which the direction of the effect size can be predicted, the novelty effect can either increase or decrease the effect size. For example, a novel intervention may increase interest levels and, consequently, motivation and participation levels, which, in turn, may be accompanied by increases in levels of performance. This sequence of events would tend to *increase* the effect size pertaining to the intervention effect. On the other hand, if a novel stimuli is introduced into the environment that is not part of the intervention but is used to collect data (e.g., a video camera), then participants can become distracted, thereby reducing their performance levels. This latter example would *reduce* the effect size. Encouragingly, the novelty effect often can be minimized by conducting the study for a period of time sufficient to allow the novelty of the intervention to subside.

Finally, the placebo effect, a term borrowed from the medical field, represents a psychological effect, in which individuals in the control group attain more favorable outcomes (e.g., more positive attitudes, higher performance levels) simply because they believed that they were in the intervention group. This phenomenon not only has the effect of reducing the effect size but negating it, and, thus, seriously affects internal validity.

Treatment diffusion. Treatment diffusion, also known as the *seepage effect*, occurs when different intervention groups communicate with each other, such that some of the treatment *seeps* out into another intervention group. Interest in each other's treatments may lead to groups borrowing aspects from each other so that the study no longer has two

or more distinctly different interventions, but overlapping interventions. In other words, the interventions are no longer independent among groups, and the integrity of each treatment is diffused. Treatment diffusion is quite common in the school setting where siblings may be in different classes and, consequently, in different intervention groups. Typically, it is the more desirable intervention that seeps out, or is diffused, into the other conditions. In this case, treatment diffusion leads to a *protocol bias* for the control groups. Thus, treatment diffusion has a tendency to reduce the effect size. However, treatment diffusion can be minimized by having strict intervention protocols and then monitoring the implementation of the interventions.

Time x treatment interaction. A time by treatment interaction occurs if individuals in one group are exposed to an intervention for a longer period of time than are individuals receiving another intervention in such a way that this differentially affects group members' responses to the intervention. Alternatively, although participants in different groups may receive their respective intervention for the same period of time, a threat to validity may prevail if one of these interventions needs a longer period of time for any positive effects to be realized. For example, suppose that a researcher wanted to compare the academic performance of students experiencing a 4x4-block scheduling model, in which students take four subjects for 90 minutes per day for the duration of a semester, to a block-8 scheduling model, in which students take the first four subjects for two days, the other four subjects for another two days, and all eight subjects on the fifth day of the week. Thus, students in the 4x4-block scheduling model are exposed to four subjects per semester for a total of eight subjects per year, whereas students in the block-8 scheduling model are

taught eight subjects per semester. If the researcher was to compare the academic performance after six months, although students in both groups would have experienced the interventions for the same period of time, a time by treatment interaction threat to internal validity likely would prevail inasmuch as students in the 4x4-block scheduling would have received more exposure to four subjects but less exposure to the other four subjects.

Another way in which time by treatment interaction can affect internal validity pertains to the amount of time that elapses between administration of the pretest and posttest. Specifically, an intervention effect based on the administration of a posttest immediately following the end of the intervention phase may not yield the same effect if a delayed posttest is administered some time after the end of the intervention phase.

History x treatment interaction. a history by treatment interaction occurs if the interventions being compared experience different history events, and that these events differentially affect group members' responses to the intervention. For example, suppose a new intervention is being compared to an existing one. However, if during the course of the study, another innovation is introduced to the school(s) receiving the new intervention, it would be impossible to separate the effects of the new intervention from the effects of the subsequent innovation. Unfortunately, it is common for schools to be exposed to additional interventions while one intervention is taking place. The difference between this particular component of history by treatment interaction threat to internal validity and the multiple-interference threat is that, whereas the researcher has no control over the former, the latter (i.e., multiple-treatment interference threat) is a function of the research design.

Threats to External Validity

The following 12 threats to external validity occur at the research design/data collection stage.

Population validity. Population validity refers to the extent to which findings are generalizable from the sample of individuals on which a study was conducted to the larger target population of individuals, as well as across different subpopulations within the larger target population. Utilizing large and random samples tend to increase the population validity of results. Unfortunately, population validity is a threat in virtually all educational studies because (a) *all* members of the target population rarely are available for selection in a study, and (b) random samples are difficult to obtain due to practical considerations such as time, money, resources, and logistics. With respect to the first consideration, most researchers are forced to select a sample from the accessible population representing the group of participants who are available for participation in the inquiry. Unfortunately, it cannot be assumed that the accessible population is representative of the target population. The degree of representativeness depends on how large the accessible population is relative to the target population. With respect to the second consideration, even if a random sample is taken, this does not guarantee that the sample will be representativeness of either the accessible or the target population. As such population validity is a threat in nearly all studies, necessitating external replications, regardless of the level of internal validity attained in a particular study.

Ecological validity. Ecological validity refers to the extent to which findings from a study can be generalized across settings, conditions, variables, and contexts. For example,

if findings can be generalized from one school to another, from one school district to another school district, or from one state to another, then the study possesses ecological validity. As such, ecological validity represents the extent to which findings from a study are independent of the setting or location in which the investigation took place. Because schools and school districts often differ substantially with respect to variables such as ethnicity, socioeconomic status, and academic achievement, ecological validity is a threat in most studies.

Temporal validity. Temporal validity refers to the extent to which research findings can be generalized across time. In other words, temporal validity pertains to the extent that results are invariant across time. Although temporal validity is rarely discussed as a threat to external validity by educational researchers, it is a common threat in the educational context because most studies are conducted at one period of time (e.g., cross-sectional studies). Thus, failure to consider the role of time at the research design/data collection stage can threaten the external validity of the study.

Multiple-treatment interference. As noted above, multiple-treatment interference occurs when the same research participants are exposed to more than one intervention. Multiple treatment interference also may occur when individuals who have already participated in a study are selected for inclusion in another, seemingly unrelated, study. It is a threat to external validity inasmuch as it is a sequencing effect that reduces a researcher's ability to generalize findings to the accessible or target population because generalization typically is limited to the particular sequence of interventions that was administered.

Researcher bias. Researcher bias, also known as experimenter effect, has been defined above in the threats to internal validity section. The reason why researcher bias also poses a threat to external validity is because the findings may be dependent, in part, on the characteristics and values of the researcher. The more unique the researcher's characteristic and values that interfere with the data collected, the less generalizable the findings.

Reactive Arrangements. Reactive arrangements, as described above in the section on internal validity, is more traditionally viewed as a threat to external validity. The five components of reactive arrangements reduce external validity because, in their presence, findings pertaining to the intervention become a function of which of these components prevail. Thus, it is not clear whether, for example, an intervention effect in the presence of the novelty effect would be the same if the novelty effect had not prevailed, thereby threatening the generalizability of the results.

Order bias. As is the case for reactive arrangements, order bias is a threat to external validity because in its presence, observed findings would depend on the order in which the multiple interventions are administered. As such, findings resulting from a particular order of administration could not be confidently generalized to situations in which the sequence of interventions is different.

Matching bias. Matching bias is a threat to external validity to the extent that findings from the matched participants could not be generalized to the results that would have occurred among individuals in the accessible population for whom a match could not be found--that is, those in the sampling frame who were not selected for the study.

Specificity of variables. Specificity of variables is a threat to external validity in almost every study. Specificity of variables refers to the fact that any given inquiry is undertaken utilizing (a) a specific type of individual; (b) at a specific time, (c) at a specific location, (d) under a specific set of circumstances, (e) based on a specific operational definition of the independent variable, (f) using specific dependent variables, and (g) using specific instruments to measure all the variables. The more unique the participants, time, context, conditions, and variables, the less generalizable the findings will be. In order to counter threats to external validity associated with specificity of variables, the researcher must operationally define variables in a way that has meaning outside of the study setting and exercise extreme caution in generalizing findings.

Treatment diffusion. Treatment diffusion threatens external validity inasmuch as the extent to which the intervention is diffused to other treatment conditions threatens the researcher's ability to generalize the findings. Like for internal validity, treatment diffusion can threaten external validity by contaminating one of the treatment conditions in a unique way that cannot be replicated.

Pretest x treatment interaction. Pretest by treatment interaction refers to situations in which the administration of a pretest increases or decreases the participants' responsiveness or sensitivity to the intervention, thereby rendering the observed findings of the pretested group unrepresentative of the effects of the independent variable for the unpretested population from which the study participants were selected. In this case, a researcher can generalize the findings to pretested groups but not to unpretested groups. The seriousness of the pretest by treatment interaction threat to external validity is

dependent on the characteristics of the research participants, the duration of the study, and the nature of the independent and dependent variables. For example, the shorter the study, the more the pre-intervention measures may influence the participants' post-intervention responses. Additionally, research utilizing self-report measures such as attitudinal scales are more susceptible to the pretest by treatment threat.

Selection x treatment interaction. Selection by treatment interaction is similar to the differential selection of participants threat to internal validity inasmuch as it stems from important pre-intervention differences between intervention groups, differences that emerge because the intervention groups are not representative of the same underlying population. Thus, it would not be possible to generalize the results from one group to another group. Although selection-treatment interaction tends to be more common when participants are not randomized to intervention groups, this threat to external validity still prevails when randomization takes place. This is because randomization does not render the group representative of the target population.

Data Analysis

Threats to Internal Validity

As illustrated in Figure 1, the following 21 threats to internal validity occur at the data analysis stage.

Statistical regression. As noted by Campbell and Kenny (1999), statistical regression can occur at the data analysis stage when researchers attempt to statistically equate groups, analyze change scores, or analyze longitudinal data. Most comparisons made in educational research involve intact groups that may have pre-existing differences.

Unfortunately, these differences often threaten the internal validity of the findings (Gay & Airasian, 2000). Thus, in an attempt to minimize this threat, some analysts utilize analysis of covariance (ANCOVA) techniques that attempt to control statistically for pre-existing differences between the groups being studied (Onwuegbuzie & Daniel, 2000). Unfortunately, most of these published works have inappropriately used ANCOVA because one or more of the assumptions have either not been checked or met (Glass, Peckham, & Sanders, 1972). According to Campbell and Kenny (1999), covariates always have measurement error, which if large, leads to a regression artifact. Further, it is virtually impossible to measure and to control for all influential covariates. For example, in comparing Black and White students, many analysts attempt to adjust for socioeconomic status or other covariates. However, almost in every case, such an adjustment represents an under-adjustment. As illustrated by Campbell and Kenny (1999), White students generally score higher on covariates than do Black students. Because these covariates are positively correlated with many educational outcomes (e.g., academic achievement), controlling for these covariates only partially adjusts for ethnic differences. Additionally, when making such comparisons, scores of each group typically regress to different means. Thus, statistical equating predicts more regression toward the mean than actually occurs (Lund, 1989).

For compensatory programs, in which the control group(s) outscore the intervention group(s) on pre-intervention measures, the bias resulting from statistical equating tends to lead to negative bias. Conversely, for anticompensatory programs, whereby intervention participants outscore the control participants, statistical equating tends to produce positive

bias (Campbell & Kenny, 1999). As such, statistical regression may mask the benefits of an effective program. Conversely, negative effects of a program can become obscured as a result of statistical regression. Simply put, statistical equating is unlikely to produce unbiased estimates of the intervention effect. Thus, researchers should be cognizant of this potential for bias when performing statistical adjustments.

Onwuegbuzie and Daniel (2000) discussed other problems associated with use of ANCOVA techniques. In particular, they note the importance of the homogeneity of regression slopes assumption. According to these theorists, to the extent that the individual regression slopes are different, the part correlation of the covariate-adjusted dependent variable with the independent variable will more closely mirror a partial correlation, and the pooled regression slope will not provide an adequate representation of some or all of the groups. In this case, the ANCOVA will introduce bias into the data instead of providing a "correction" for the confounding variable (Loftin & Madison, 1991). Ironically, as noted by Henson (1998), ANCOVA typically is appropriate when used with randomly assigned groups; however, it is typically not justified when groups are not randomly assigned.

Another argument against the use of ANCOVA is that after using a covariate to adjust the dependent variable, it is not clear whether the residual scores are interpretable (Thompson, 1992). Disturbingly, some researchers utilize ANCOVA as a substitute for not incorporating a true experimental design, believing that methodological designs and statistical analyses are synonymous (Henson, 1998; Thompson, 1994b). In many cases, statistical equating creates the illusion of equivalence but not the reality. Indeed, the problems with statistical adjusting has prompted Campbell and Kenny to declare: "The

failure to understand the likely direction of bias when statistical equating is used is one of the most serious difficulties in contemporary data analysis" (p. 85).

A popular statistical technique is to measure the effect of an intervention by comparing pre-intervention and post-intervention scores, using analyses such as the dependent (matched-pairs) *t*-test. Unfortunately, this type of analysis is affected by regression to the mean, which tends to reduce the effect size (Campbell & Kenny, 1999). Also, as stated by Campbell and Kenny (1999), "in longitudinal studies with many periodic waves of measurement, anchoring the analysis...at any one time...is likely to produce an ever-increasing pseudo effect as the time interval increases" (p. 139). Thus, analysis of both change scores and longitudinal data can threaten internal validity.

Restricted range. Lacking the knowledge that virtually all parametric analyses represent the general linear model, many researchers inappropriately categorize variables in non-experimental designs using ANOVA, in an attempt to justify making causal inferences, when all that occurs typically is a discarding of relevant variance (Cliff, 1987; Onwuegbuzie & Daniel, 2000; Pedhazur, 1982; Prosser, 1990). For example, Cohen (1983) calculated that the Pearson product-moment correlation between a variable and its dichotomized version (i.e., divided at the mean) was .798, which suggests that the cost of dichotomization is approximately a 20% reduction in correlation coefficient. In other words, an artificially dichotomized variable accounts for only 63.7% (i.e., $.798^2$) as much variance as does the original continuous variable. It follows that with factorial ANOVAs, when artificial categorization occurs, even more power is sacrificed. Thus, restricting the range of scores by categorizing data tends to pose a threat to internal validity at the data analysis

stage by reducing the size of the effect.

Thus, as stated by Kerlinger (1986), researchers should avoid artificially categorizing continuous variables, unless compelled to do so as a result of the distribution of the data (e.g., bimodal). Indeed, rather than categorizing independent variables, in many cases, regression techniques should be used, because they have been shown consistently to be superior to OVA methods (Onwuegbuzie & Daniel, 2000).

Mortality. In an attempt to analyze groups with equal or approximately equal sample sizes (i.e., to undertake a "balanced" analysis), some researchers remove some of the participants' scores from their final dataset. That is, the size of the largest group(s) is deliberately reduced to resemble more closely the size of the smaller group(s). Whether or not cases are removed randomly, this practice poses a threat to internal validity the extent to which the participants who are removed from the dataset are different than those who remain. That is, the practice of sub-sampling from a dataset introduces or adds bias into the analysis, influencing the effect size in an unknown manner.

Non-Interaction seeking bias. Many researchers neglect to assess the presence interactions when testing hypotheses. By not formally testing for interactions, researchers may be utilizing a model that does not honor, in the optimal sense, the nature of reality that they want to study, thereby providing a threat to internal validity at the data analysis stage.

Type I to Type X error. Daniel and Onwuegbuzie (2000) have identified 10 errors associated with statistical significance testing. These errors were labeled Type I to Type X. The first four errors are known to all statisticians as Type I (falsely rejecting the null hypothesis), Type II (incorrectly failing to reject the null hypothesis), Type III (incorrect

inferences about result directionality), and Type IV (incorrectly following-up an interaction effect with a simple effects analysis). The following six additional types of error were identified by Daniel and Onwuegbuzie (2000): (a) Type V error--internal replication error–measured via incidence of Type I or Type II errors detected during internal replication cycles when using methodologies such as the jackknife procedure; (b) Type VI error--reliability generalization error--measured via linkages of statistical results to characteristics of scores on the measures used to generate results (a particularly problematic type of error when researchers fail to consider differential reliability estimates for subsamples within a data set); (c) Type VII error—heterogeneity of variance/regression—measured via the extent to which data treated via analysis of variance/covariance are not appropriately screened to determine whether they meet homogeneity assumptions prior to analysis of group comparison statistics; (d) Type VIII error—test directionality error—measured as the extent to which researchers express alternative hypotheses as directional yet assess results with two-tailed tests; (e) Type IX error—sampling bias error—measured via disparities in results generated from numerous convenience samples across a multiplicity of similar studies; and (f) Type X error—degrees of freedom error—measured as the tendency of researchers using certain statistical procedures (chiefly stepwise procedures) erroneously to compute the degrees of freedom utilized in these procedures. All of these errors pose a threat to internal validity at the data analysis stage.

Observational bias. In studies when observations are made, an initial part of the data analysis often involves coding the observations. Whenever inter-rater reliability of the

coding scheme is less than 100%, internal validity is threatened. Thus, researchers should always attempt to assess the inter-rater reliability of any coding of observations. When inter-reliability estimates cannot be obtained because there is only one rater, intra-rater reliability estimates should be assessed.

Researcher bias. Perhaps the biggest form of researcher bias is what has been termed the *halo effect*. The halo effect occurs when a researcher is evaluating open-ended responses, or the like, and allows his or her prior knowledge of the participants to influence the scores given. This results in findings that are biased. Clearly, this is a threat to internal validity at the data analysis stage.

Matching bias. Another common data analysis technique is to match groups after the data on the complete sample have been collected. Unfortunately, the ability of matching to equate groups, again often is more of an illusion than a reality (Campbell & Kenny, 1999). Moreover, bias is introduced as a result of omitting those who were not matched, providing a threat to internal validity.

Treatment replication error. Using an inappropriate unit of analysis is a common mistake made by researchers (McMillan, 1999). The treatment replication error threat to internal validity occurs at the data analysis stage when researchers utilize an incorrect unit of analysis even though data are available for them to engage in a more appropriate analysis. For example, in analyzing data pertaining to cooperative learning groups, an investigator may refrain from analyzing available group scores. That is, even though the intervention is given to groups of students, the researcher might incorrectly use individual students as the unit of analysis, instead of utilizing each group as a treatment unit and

analyzing group data. Unfortunately, analyzing individual students' scores does not take into account possible confounding factors. Although it is likely that analyzing group data instead of individual data results in a loss of statistical power due to a reduction in the number of treatment units, the loss in power typically is compensated for by the fact that using group data is more free from contamination. Moreover, analyzing individual data when groups received the intervention violates the independence assumption, thereby providing a serious threat to internal validity. Usually, independence violations tend to inflate both the Type I error rate and effect size estimates. Thus, researchers always should analyze data at the group level for subsequent analysis when there is a limited number of interventions independently replicated.

Violated assumptions. Disturbingly, it is clear that many researchers do not adequately check the underlying assumptions associated with a particular statistical test. This is evidenced by the paucity of researchers who provide information about the extent to which assumptions are met (see for example, Keselman et al., 1998; Onwuegbuzie, 1999). Thus, researchers always should check model assumptions. For example, if the normality assumption is violated, analysts should utilize the non-parametric counterparts.

Multicollinearity. Most analysts do not appear to evaluate multicollinearity among the regression variables (Onwuegbuzie & Daniel, 2000). However, multicollinearity is a more common threat than researchers acknowledge or appear to realize. For example, race/ethnicity and socioeconomic status often are confounded with each other in such a way that the presence of one variable in a model may affect the predictive power of the other variable. Moreover, multicollinearity leads to inflated or unstable statistical

coefficients, thereby providing rival explanations for the findings. Thus, multicollinearity should routinely be assessed in multiple regression models.

Mis-specification error. Mis-specification error is perhaps the most hidden threat to internal validity. This error, which involves omitting one or more important variables from the final model, often stems from a weak or non-existent theoretical framework for building a statistical model. This inattention to a theoretical framework leads many researchers to utilize data-driven techniques such as stepwise multiple regression procedures (i.e., forward selection, backward selection, stepwise selection). Indeed, the use of stepwise regression in educational research is rampant (Huberty, 1994), probably due to its widespread availability on statistical computer software programs. As a result of this seeming obsession with stepwise regression, as stated by Cliff (1987, pp. 120-121), "a large proportion of the published results using this method probably present conclusions that are not supported by the data."

Mis-specification error also includes *non-interaction seeking bias*, discussed above, in which interactions are not tested. Indeed, this is a particular problem when undertaking structural equation modeling (SEM) techniques. Many SEM software do not facilitate the statistical testing of interaction terms. Unfortunately, mis-specification error, although likely common, is extremely difficult to detect, especially if the selected *non-optimal model*, which does not include any interaction terms, fits the data adequately.

Threats to External Validity

As illustrated in Figure 1, the following five threats to external validity occur at the data analysis stage: population validity, researcher bias, specificity of variables, matching

bias, and mis-specification error. All of these threats have been discussed above. Thus, only a brief mention will be made of each.

Population validity. Every time a researcher analyzes a subset of her or his dataset, it is likely that findings emerging from this subset are less generalizable than are those that would have arisen if the total sample had been used. In other words, any kind of sub-sampling from the dataset likely decreases population validity. The greater the discrepancy between those sampled and those not sampled from the full dataset, the greater the threat to population validity. Additionally, threats to population validity often occur at the data analysis stage because researchers fail to disaggregate their data, incorrectly assuming that their findings are invariant across all sub-samples inherent in their study. In fact, when possible, researchers should utilize *condition-seeking* methods, whereby they "seek to discover which, of the many conditions that were confounded together in procedures that have obtained a finding, are indeed necessary or sufficient" (Greenwald et al., 1986, p. 223).

Researcher bias. Researcher bias, such as the halo effect, not only affects internal validity at the data analysis stage, but also threatens external validity because the particular type of bias of the researcher may be so unique as to make the findings ungeneralizable.

Specificity of variables. As noted above, specificity of variables is one of the most common threats to external validity at the research design/data collection stage. Indeed, seven ways in which specificity of variables is a threat to external validity at this stage was identified above (type of participants, time, location, circumstance, operational definition

of the independent variables, operational definition of the dependent variables, and types of instruments used). At the data analysis stage, specificity of variables also can be an external validity threat vis-à-vis the manner in which the independent and dependent variables are operationalized. For example, in categorizing independent and dependent variables, many researchers use local norms; that is, they classify participants' scores based on the underlying distribution. Because every distribution of scores is sample specific, the extent to which a variable categorized using local norms can be generalized outside the sample is questionable. Simply put, the more unique the operationalization of the variables, the less generalizable will be the findings. In order to counter threats to external validity associated operationalization of variables, when possible, the researcher should utilize variables in ways that are transferable (e.g., using standardized norms).

Matching bias. Some researchers match individuals in the different intervention groups just prior to analyzing the data. Matching provides a threat to external validity at this stage if those not selected for matching from the dataset are in some important way different than those who are matched, such that the findings from the selected individuals may not be generalizable to the unselected persons.

Mis-specification error. As discussed above, mis-specification error involves omitting one or more important variables (e.g., interaction terms) from the analysis. Although a final model selected may have acceptable internal validity, such omission reduces the external validity because it is not clear whether the findings would be the same if the omitted variable(s) had been included.

Data Interpretation

Threats to Internal Validity

As illustrated in Figure 1, the following seven threats to internal validity occur at the data interpretation stage.

Effect size. As noted by Onwuegbuzie and Daniel (2000), perhaps the most prevalent error made in quantitative research, which appears across all types of inferential analyses, involves the incorrect interpretation of statistical significance and the related failure to report and to interpret confidence intervals and effect sizes (i.e., variance-accounted for effect sizes or standardized mean differences) (Daniel, 1998a, 1998b; Ernest & McLean, 1998; Knapp, 1998; Levin, 1998; McLean & Ernest, 1998; Nix & Barnette, 1998a, 1998b; Thompson, 1998b). This error, which occurs at the data interpretation stage, threatens internal validity because it often leads to under-interpretation of associated *p*-values when sample sizes are small and the corresponding effect sizes are large, and an over-interpretation of *p*-values when sample sizes are large and effect sizes are small (e.g., Daniel, 1998a). Because of this common confusion between significance in the probabilistic sense (i.e., statistical significance) and significance in the practical sense (i.e., effect size), some researchers (e.g., Daniel, 1998a) have recommended that authors insert the word "statistically" before the word "significant," when interpreting the findings of a null hypothesis statistical test. Thus, as stated by the APA Task Force, researchers should "always present effect sizes for primary outcomes...[and]...reporting and interpreting effect sizes...is essential to good research" (Wilkinson & the Task Force on Statistical Inference, 1999, pp. 10-11).

Confirmation bias. Confirmation bias is the tendency for interpretations and

conclusions based on new data to be overly consistent with preliminary hypotheses (Greenwald et al., 1986). Unfortunately, confirmation bias is a common threat to external validity at the data interpretation stage, and has been identified via expectancy biasing of student achievement, perseverance of belief in discredited hypotheses, the primacy effects in impression formation and persuasion, delayed recovery of simple solutions, and selective retrieval of information that confirms the researcher's hypotheses, opinions, or self-concept (Greenwald et al., 1986). Apparently, confirmation bias is more likely to prevail when the researcher is seeking to test theory than when he or she is attempting to generate theory, because testing a theory can "dominate research in a way that blinds the researcher to potentially informative observation" (Greenwald et al., 1986, p. 217). When hypotheses are not supported, a common practice of researchers is to proceed as if the theory underlying the hypotheses is still likely to be correct. In proceeding in this manner, many researchers fail to realize that their research methodology no longer can be described as theory testing but theory confirming.

Notwithstanding, confirmation bias, per se, does not necessarily pose a threat to internal validity. It threatens internal validity at the data interpretation stage only when there exists one or more plausible rival explanations to underlying findings that might be demonstrated to be superior if given the opportunity. Conversely, when no rival explanations prevail, confirmation bias helps to provide support for the best or sole explanation of results (Greenwald et al., 1986). However, because rival explanations typically permeate educational research studies, researchers should be especially cognizant of the role of confirmation bias on the internal validity of the results at the data

interpretation stage.

Statistical regression. When a study involves extreme group selection, matching, statistical equating, change scores, time-series studies, or longitudinal studies, researchers should be especially careful when interpreting data because, as noted above, findings from such investigation often reflect some degree of regression toward the mean (Campbell & Kenny, 1999).

Distorted graphics. Researchers should be especially careful when interpreting graphs. In particular, when utilizing graphs (e.g., histograms) to check model assumptions, in a desire to utilize parametric techniques, it is not unusual for researchers to conclude incorrectly that these assumptions hold. Thus, when possible graphical checks should be triangulated by empirical evaluation. For example, in addition to examining histograms, analysts could examine the skewness and kurtosis coefficients, and even undertake statistical tests of normality (e.g., the Shapiro-Wilk test; Shapiro & Wilk, 1965; Shapiro, Wilk, & Chen, 1968).

Illusory correlation. The illusory correlation represents a tendency to overestimate the relationship among variables that are only slightly related or not related at all. Often, the illusory correlation stems from a confirmation bias. The illusory correlation also may arise from a *false consensus bias*, in which researchers have the false belief that most other individuals share their interpretations of a relationship. Such an illusory correlation poses a serious threat to internal validity at the data interpretation stage.

Crud factor. As noted by Onwuegbuzie and Daniel (in press), as the sample size increases, so does the probability of rejecting the null hypothesis of no relationship

between two variables. Indeed, theoretically, given a large enough sample size, the null hypothesis always will be rejected (Cohen, 1994). Hence, it can be argued that "everything correlates to some extent with everything else" (Meehl, 1990, p. 204). Meehl referred to this tendency to reject null hypotheses in the face of trivial relationships as the *crud* factor. This crud factor leads some researchers to identify and to interpret relationships that are not real but represent statistical artifacts, posing a threat to internal validity at the data interpretation stage.

Positive manifold. Positive manifold refers to the phenomenon that individuals who perform well on one ability or attitudinal measure tend to perform well on other measures in the same domain (Neisser, 1998). Thus, researchers should be careful when interpreting relationships found between two or more sets of cognitive test scores or attitudinal scores. As noted by Onwuegbuzie and Daniel (in press), particular focus should be directed toward effect sizes, as opposed to p-values.

Causal error. In interpreting statistically significant relationships, often infer cause-and-effect relationships, even though such associations can, at best, only be determined from experimental studies. Causality often can be inferred from scientific experiments when the selected independent variable(s) are carefully controlled. Then if the dependent variable is observed to change in a predictable way as the value of the independent variable changes, the most plausible explanation would be a causal relationship between the independent and the dependent variables. In the absence of such control and ability to manipulate the independent variable, the plausibility that at least one more unidentified variable is mediating the relationship between both variables will remain.

Interestingly, Kenny (1979) distinguished between correlational and causal inferences, noting that four conditions must exist before a researcher may justifiably claim that *X* causes *Y*: (a) time precedence (*X* must precede *Y* in time), (b) functional relationship (*Y* should be conditionally distributed across *X*), (c) nonspuriousness (there must not be a third variable *Z* that causes both *X* and *Y*, such that when *Z* is controlled for, the relationship between *X* and *Y* vanishes, and (d) vitality (a logistical link between *X* and *Y* that substantiates the likelihood of a causal link (such as would be established via controlled experimental conditions). However, it is extremely difficult for these four conditions to be met simultaneously in correlational designs. Consequently, substantiating causal links in uncontrolled (correlational and intervention) studies is a very difficult task (Onwuegbuzie & Daniel, in press). Thus, researchers should pay special attention when interpreting findings stemming from non-experimental research. Unfortunately, some researchers and policy makers are prone to ignore threats to internal validity when interpreting relationships among variables.

Threats to External Validity

As illustrated in Figure 1, the following three threats to external validity occur at the data interpretation stage: population validity, ecological validity, and temporal validity. All of these threats have been discussed above. Thus, only a brief mention will be made of each.

Population validity/Ecological validity/Temporal validity. When interpreting findings stemming from small and/or non-random samples, researchers should be very careful not to over-generalize their conclusions. Instead, researchers always should compare their

findings to the extant literature as comprehensively as is possible, so that their results can be placed in a realistic context. Only if findings are consistent across different populations, locations, settings, times, and contexts can researchers be justified in making generalizations to the target population. Indeed, researchers and practitioners must refrain from assuming that one study, conducted without any external replications, can ever adequately answer a research question. Thus, researchers should focus more on advocating external replications and on providing directions for future research than on making definitive conclusions. When interpreting findings, researchers should attempt to do so via the use of disaggregated data, utilizing the *condition-seeking* methods, in which a progression of qualifying conditions are made based on existing findings (Greenwald et al., 1986). Such condition-seeking methods would generate a progression of research questions, which, if addressed in future studies, would provide increasingly accurate and generalizable conclusions. Simply put, researchers should attempt, at best, to make qualified conclusions.

Summary and Conclusions

The present paper has sought to promote the dialogue about threats to internal and external validity in educational research in general and empirical research in particular. First, several rationales were provided for identifying and discussing threats to internal and external validity not only in experimental studies, but for all other types quantitative research designs (e.g., descriptive, correlational, causal-comparative). Specifically, it was contended that providing information about sources of invalidity and rival explanations (a) allows readers better to contextualize the underlying findings, (b) promotes external

replications; (c) provides a directions for future research, and (d) advances the conducting of validity meta analyses and thematic effect sizes.

Second, the validity frameworks of Campbell and Stanley (1963), Huck and Sandler (1979), and McMillan (2000) were described. It was noted that these three sets of theorists are the only ones who appear to have provided a list of internal and external validity threats. Unfortunately, none of these frameworks identified sources of result invalidity that were applicable across all types of quantitative research designs. Third, it was asserted that in order to encourage empirical discussion of internal and external validity threats in *all* empirical studies, a framework was needed that is more comprehensive than are the existing ones, and that seeks to unify all quantitative research designs under one validity umbrella.

Fourth, threats to internal and external validity were conceptualized as occurring at the three major stages of the research process, namely, research design/data collection, data analysis, and data interpretation. Using this conceptualization, and building on the works of Campbell and Stanley (1963), Huck and Sandler (1979), and McMillan (2000), a comprehensive model of dimensions of sources of validity was developed. This model was represented as a 3 (stage of research process) x 2 (internal vs. external validity) matrix comprising 49 unique dimensions of internal and external validity threats, with many of the dimensions containing sub-dimensions (cf. Figure 1).

Although this model of sources of validity is comprehensive, it is by no means exhaustive. Indeed, researchers and practitioners alike are encouraged to find ways to improve upon this framework. Indeed, the author currently is formally assessing the internal

and external validity of this model by attempting to determine how prevalent each of these threats are in the extant educational literature.

Nevertheless, it is hoped that this paper makes it clear that *every* inquiry contains multiple threats to internal and external validity, and that researchers should exercise extreme caution when making conclusions based on one or a few studies. Additionally, it is hoped tat this model highlights the importance of assessing sources of invalidity in every research study and at different stages of the research process. For example, just because threats to internal and external validity have been minimized at one phase of the research study does not mean that sources of invalidity do not prevail at the other stages.

Moreover, it is hoped that the present model not only extends the dialogue on threats to internal and external validity, but also provides a broader guideline for doing so than has previously been undertaken. However, in order to promote further discussion of these threats, journal editors must be receptive to this information, and not use it as a vehicle to justify the rejection of manuscripts. Indeed, journal reviewers and editors should strongly encourage all researchers to include a discussion of the major rival hypotheses in their investigations. In order to motivate researchers to undertake this, it must be made clear to them that such practice would improve the quality of their papers, not diminish it. Indeed, future revisions of the *American Psychological Association Publication Manual* (APA, 1994) should provide *strong* encouragement for *all* empirical research reports to include a discussion of threats to internal and external validity. Additionally, the Manual should urge researchers to furnish a summary of the major threats to internal and external validity for some or even all of the studies that are included in their reviews of the related

literature. Unless there is a greater emphasis on validity in research, threats to internal and external validity will continue to prevail at various stages of the research design, and many findings will continue to be misinterpreted and over-generalized. Thus, an increased focus on internal and external validity in all empirical studies can only help the field of educational research by helping investigators to be more reflective at every stage of the research process.

References

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297-312.

Campbell, D.T., & Kenny, D.A. (1999). *a primer on regression artifacts*. New York: The Guildford Press.

Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cliff, N. (1987). *Analyzing multivariate data*. San Diego: Harcourt Brace Jovanovich.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249-253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: John Wiley.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Daniel, L.G. (1998a). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for editorial policies of educational journals. *Research in the Schools, 5*, 23-32.

Daniel, L.G. (1998b). The statistical significance controversy is definitely not over: A rejoinder to responses by Thompson, Knapp, and Levin. *Research in the Schools, 5*, 63-65.

Daniel, L.G., & Onwuegbuzie, A.J. (2000, November). *Toward an extended typology of research errors*. Paper presented at the annual conference of the Mid-South Educational Research Association, Bowling Green, KY.

Ernest, J.M., & McLean, J.E. (1998). Fight the good fight: A response to Thompson, Knapp, and Levin. *Research in the Schools, 5*, 59-62.

Gay, L.R., & Airasian, P.W. (2000). *Educational research: Competencies for analysis and application* (6th ed.). Englewood Cliffs, N.J.: Prentice Hall.

Glass, G.V., Peckham, P.D., & Sanders, J.R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*, 237-288.

Greenwald, A.G., Pratkanis, A.R., Leippe, M.R., & Baumgardner, M.H. (1986). Under what conditions does theory obstruct research progress. *Psychological Review, 93*, 216-229.

Henson, R.K. (1998, November). *ANCOVA with intact groups: Don't do it!* Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, LA.

Hill, K.T., & Wigfield, A. (1984). Test anxiety: A major educational problem and what can be done about it. *The Elementary School Journal, 85*, 105-126.

Huberty, C.J. (1994). *Applied discriminant analysis*. New York: Wiley and Sons.

Huck, S.W., & Sandler, H.M. (1979). *Rival hypotheses: Alternative interpretations of data based conclusions*. New York: Harper Collins.

Johnson, B., & Christensen, L. (2000). *Educational research: Quantitative and*

qualitative approaches. Boston, MA: Allyn and Bacon.

Kenny, D. A. (1979). *Correlation and causality*. New York: John Wiley & Sons.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York: Holt, Rinehart and Winston.

Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R.A., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C., & Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386.

Knapp, T.R. (1998). Comments on the statistical significance testing articles. *Research in the Schools, 5*, 39-42.

Levin, J.R. (1998). What if there were no more bickering about statistical significance tests? *Research in the Schools, 5*, 43-54.

Lincoln, Y.S., & Guba, E.G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Loftin, L.B., & Madison, S.Q. (1991). The extreme dangers of covariance corrections. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 133-147). Greenwich, CT: JAI Press.

Lund, T. (1989). The statistical regression phenomenon: II. Application of a metamodel. *Scandinavian Journal of Psychology, 30*, 2-11.

McLean, J.E., & Ernest, J.M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5*, 15-22.

McMillan, J.H. (1999). Unit of analysis in field experiments: Some design considerations for educational researchers. (ERIC Document Reproduction Service No.

ED 428 135)

McMillan, J.H. (2000, April). *Examining categories of rival hypotheses for educational research*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Meehl, P. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195-244.

Miles, M.B., & Huberman, A.M. (1984). Drawing valid meaning from qualitative data: Toward a shared craft. Educational Researcher, 13, 20-30.

Mundfrom, D.J., Shaw, D.G., Thomas, A., Young, S., & Moore, A.D. (1998, April). *Introductory graduate research courses: An examination of the knowledge base*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Neisser, U. (1998). Rising test scores. In U. Neisser (Ed.), *The rising curve* (pp. 3-22). Washington, DC: American Psychological Association.

Nix, T.W., & Barnette, J. J. (1998a). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5*, 3-14.

Nix, T.W., & Barnette, J. J. (1998b). A review of hypothesis testing revisited: Rejoinder to Thompson, Knapp, and Levin. *Research in the Schools, 5*, 55-58.

Onwuegbuzie, A.J. (1999, September). *Common analytical and interpretational errors in educational research*. Paper presented at the annual meeting of the European Educational Research Association (EERA), Lahti, Finland.

Onwuegbuzie, A.J. (2000a). *The prevalence of discussion of threats to internal and*

external validity in research reports. Manuscript submitted for publication.

Onwuegbuzie, A.J. (2000b, November). *Positivists, post-positivists, post-structuralists, and post-modernists: Why can't we all get along? Towards a framework for unifying research paradigms*. Paper to be presented at the annual meeting of the Association for the Advancement of Educational Research (AAER), Ponte Vedra, Florida.

Onwuegbuzie, A.J. (2000c, November). *Effect sizes in qualitative research*. Paper to be presented at the annual conference of the Mid-South Educational Research Association, Bowling Green, KY.

Onwuegbuzie, A.J., & Collins, K.M. (2000). *Group heterogeneity and performance in graduate-level educational research courses: The role of aptitude by treatment interactions and Matthew effects*. Manuscript submitted for publication.

Onwuegbuzie, A.J., & Daniel, L.G. (2000, April). *Common analytical and interpretational errors in educational research*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Onwuegbuzie, A.J., & Daniel, L.G. (in press). Uses and misuses of the correlation coefficient. *Research in the Schools*.

Onwuegbuzie, A.J., & Seaman, M. (1995). The effect of time and anxiety on statistics achievement. *Journal of Experimental Psychology, 63*, 115-124.

Pedhazur, E.J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart and Winston.

Prosser, B. (1990, January). *Beware the dangers of discarding variance*. Paper presented at the annual meeting of the Southwest Educational Research Association,

Austin, TX. (ERIC Reproduction Service No. ED 314 496)

Rogers, E.M. (1995). *Diffusion of innovations* (4th ed.). New York: The Free Press.

Shapiro, S.S., & Wilk, M.B. (1965). An analysis of variance test, for normality and complete samples. *Biometrika, 52,* 592-611.

Shapiro, S.S., Wilk, M.B., & Chen, H.J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association, 63,* 1343-1372.

Smith, M.L., & Glass, G.V. (1987). *Research and evaluation in education and the social sciences*. Englewood Cliffs, NJ: Prentice Hall.

The American Statistical Association. (1999). Ethical guidelines for statistical practice [On-line]. Available: *http://www.amstat.org/profession/ethicalstatistics.html*

Thompson, B. (1992, April). Misuse of ANCOVA and related "statistical control" procedures. *Reading Psychology: An International Quarterly, 13,* iii-xvii.

Thompson, B. (1994a). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality, 62,* 157-176.

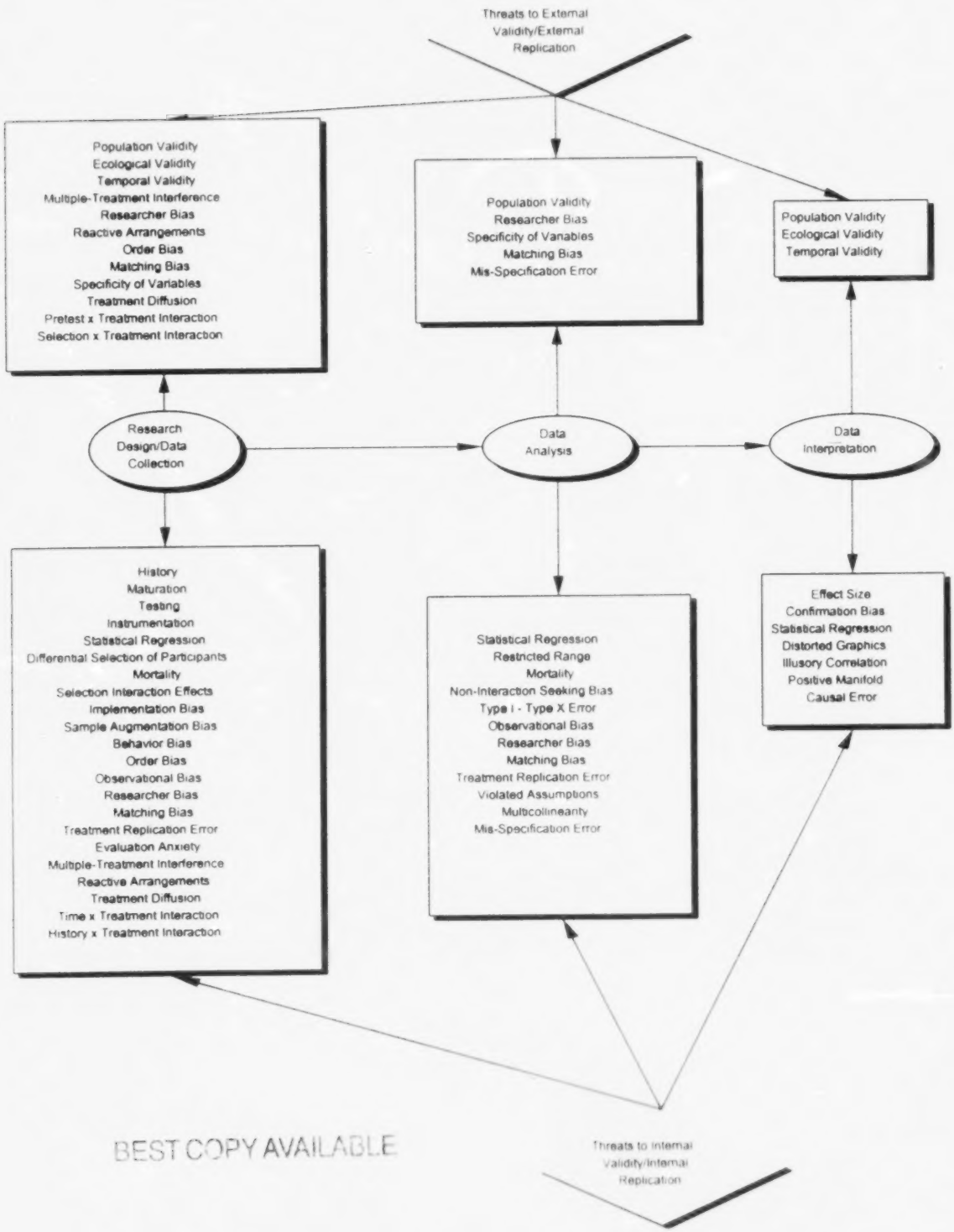
Thompson, B. (1994b). *Common methodological mistakes in dissertations, revisited*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA (ERIC Document Reproduction Service No. ED 368 771)

Thompson, B. (1998b). Statistical testing and effect size reporting: Portrait of a possible future. *Research in the Schools, 5,* 33-38.

Wilkinson, L. & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Figure Caption

Figure 1. Major dimensions of threats to internal validity and external validity at the three major stages of the research process.



BEST COPY AVAILABLE